# A Bayesian-motivated test for linear model in high-dimensional setting

Rui Wang

Thursday $15^{\text{th}}$ November, 2018

## 1 Introduction

Suppose we would like to compare models $\mathcal{M}_0$ and $\mathcal{M}_1$.

$$\begin{aligned}\mathcal{M}_0 : \mathbf{y} = \mathbf{X}_a\boldsymbol{\beta}_a + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \phi^{-1}\mathbf{I}_n),\\ \mathcal{M}_1 : \mathbf{y} = \mathbf{X}_a\boldsymbol{\beta}_a + \mathbf{X}_b\boldsymbol{\beta}_b + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \phi^{-1}\mathbf{I}_n).\end{aligned}$$

Here $\boldsymbol{\beta}_a$ is $q$ dimensional and $\boldsymbol{\beta}_b$ is $p$ dimensional. We assume that as $n$ tends to infinity, $q$ is fixed while $p/n \to \infty$. This assumption is reasonable. In practice, $p_0$ is often 1 and $\mathbf{X}_0$ is $\mathbf{1}_n$.

Although several tests have been proposed, the following proposition implies that there is no unbiased test.

**Proposition 1.** *Suppose $\mathbf{y} \sim \mathcal{N}_n(\mu, \phi^{-1}\mathbf{I}_n)$. We test $H_0 : \mu = \mathbf{X}_a\boldsymbol{\beta}_a, \boldsymbol{\beta}_a \in \mathbb{R}^q$ versus $H_1 : \mu \in \mathbb{R}^n$, where $\mathbf{X}_a$ is an $n \times q$ matrix with full column rank, $q < n$. Let $\varphi(\mathbf{y})$ be a test function, that is, a Borel measurable function, $0 \leq \phi(\mathbf{y}) \leq 1$. If $\int \varphi(\mathbf{y})\mathcal{N}_n(\mathbf{X}_a\boldsymbol{\beta}_a, \phi^{-1}\mathbf{I}_n)(d\mathbf{y}) = \alpha$ for $\boldsymbol{\beta}_a \in \mathbb{R}^q$, $\phi > 0$ and $\int \varphi(\mathbf{y})\mathcal{N}_n(\mu, \phi^{-1}\mathbf{I}_n)(d\mathbf{y}) \geq \alpha$ for $\mu \in \mathbb{R}^n$, $\phi > 0$, then $\varphi(\mathbf{y}) = \alpha$, a.s.*

So we can not find a universally good test. Instead, we would like to find a test with good average behaviour. So Bayesian methods are natural choices in this case.

Bayes hypothesis testing use the Bayes factor.

$$B_{10} = \frac{\int f_1(y|\boldsymbol{\beta}_b, \boldsymbol{\beta}_a, \phi)\pi_1(\boldsymbol{\beta}_b, \boldsymbol{\beta}_a, \phi)d\boldsymbol{\beta}_b d\boldsymbol{\beta}_a d\phi}{\int f_0(y|\boldsymbol{\beta}_a, \phi)\pi_0(\boldsymbol{\beta}_a, \phi)d\boldsymbol{\beta}_a d\phi}.$$

There have been several extensions of $g$-priors to $p > n$ case: Maruyama and George (2011), Shang and Clayton (2011).

Under $\mathcal{M}_0$, we impose the reference prior $\pi_0(\boldsymbol{\beta}_a, \phi) = c/\phi$. Note that under $\mathcal{M}_1$, the posterior corresponding to the referece prior is proper only if $n > q+p$? That is, the minimal training sample size is $q + p + 1$. So we cannot impose the reference prior under $\mathcal{M}_1$ provided $q + p + 1 > n$. We temporarily impose the conditional prior $\boldsymbol{\beta}_b|\boldsymbol{\beta}_a, \phi \sim \mathcal{N}_p(0, \kappa^{-1}\phi^{-1}\mathbf{I}_p)$. There are many literature consider the choice of $\kappa$. Kass and Wasserman (1995) choose $\kappa$ such that the amount of information about the parameter equal to the amount of information contained in one observation. Thus, under $\mathcal{M}_1$, we put prior

$$\pi_1(\boldsymbol{\beta}_b|\boldsymbol{\beta}_a,\phi) = \frac{(\kappa\phi)^{p/2}}{(2\pi)^{p/2}}\exp\left\{-\frac{\kappa\phi}{2}\|\boldsymbol{\beta}_b\|^2\right\}, \quad \pi_1(\boldsymbol{\beta}_a,\phi)=\frac{c}{\phi}.$$

It is straightforward to show that the Bayes factor associated with these priors is

$$\begin{aligned}B_{10}^{\kappa} =& \frac{\kappa^{p/2}}{|\mathbf{X}_b^\top(\mathbf{I}_n-\mathbf{P}_a)\mathbf{X}_b+\kappa\mathbf{I}_p|^{1/2}}\cdot \\ &\left(\frac{\mathbf{y}\top(\mathbf{I}_n-\mathbf{P}_a)\mathbf{y}}{\mathbf{y}\top(\mathbf{I}_n-\mathbf{P}_a)\mathbf{y}-\mathbf{y}\top(\mathbf{I}_n-\mathbf{P}_a)\mathbf{X}_b\left(\mathbf{X}_b^\top(\mathbf{I}-\mathbf{P}_a)\mathbf{X}_b+\kappa\mathbf{I}_p\right)^{-1}\mathbf{X}_b^\top(\mathbf{I}_n-\mathbf{P}_a)\mathbf{y}}\right)^{(n-q)/2}.\end{aligned}$$

Thus,

$$\begin{aligned}2\log B_{10}^{\kappa} =& p\log\kappa-\log|\mathbf{X}_b^\top(\mathbf{I}_n-\mathbf{P}_a)\mathbf{X}_b+\kappa\mathbf{I}_p| \\ &-(n-q)\log\left(1-\frac{\mathbf{y}\top(\mathbf{I}_n-\mathbf{P}_a)\mathbf{X}_b\left(\mathbf{X}_b^\top(\mathbf{I}-\mathbf{P}_a)\mathbf{X}_b+\kappa\mathbf{I}_p\right)^{-1}\mathbf{X}_b^\top(\mathbf{I}_n-\mathbf{P}_a)\mathbf{y}}{\mathbf{y}\top(\mathbf{I}_n-\mathbf{P}_a)\mathbf{y}}\right).\end{aligned}$$

Denote by $\mathbf{I}_n-\mathbf{P}_a=\tilde{\mathbf{U}}_a\tilde{\mathbf{U}}_a^\top$ the rank decomposition of $\mathbf{I}_n-\mathbf{P}_a$, where $\tilde{\mathbf{U}}_a$ is a $n\times(n-q)$ column orthogonal matrix. Let $\mathbf{X}_b^*=\tilde{\mathbf{U}}_a^\top\mathbf{X}_b$, $\mathbf{y}^*=\tilde{\mathbf{U}}_a^\top\mathbf{X}_b$. Let $\gamma_i$ be the $i$th largest eigenvalue of $\mathbf{X}_b^*\mathbf{X}_b^{*\top}$, $i=1,\ldots,n-q$. Denote by $\mathbf{X}_b^*=\mathbf{U}_b^*\mathbf{D}_b^*\mathbf{V}_b^{*\top}$ the singular value decomposition of $\mathbf{X}_b^*$, where $\mathbf{U}_b^*$, $\mathbf{V}_b^*$ are $(n-q)\times(n-q)$ and $p\times(n-q)$ column orthogonal matrices, respectively, and $\mathbf{D}_b^*=\operatorname{diag}(\sqrt{\gamma_1},\ldots,\sqrt{\gamma_{n-q}})$. Then

$$\begin{aligned}2\log B_{10}^{\kappa} =& p\log\kappa-\sum_{i=1}^{n-q}\log(\gamma_i+\kappa)-(p-(n-q))\log\kappa \\ &-(n-q)\log\left(1-\frac{\mathbf{y}^{*\top}\mathbf{X}_b^*\left(\mathbf{X}_b^{*\top}\mathbf{X}_b^*+\kappa\mathbf{I}_p\right)^{-1}\mathbf{X}_b^{*\top}\mathbf{y}^*}{\mathbf{y}^{*\top}\mathbf{y}^*}\right) \\ =& -\sum_{i=1}^{n-q}\log(\gamma_i+\kappa)+(n-q)\log\left(\frac{\mathbf{y}^{*\top}\mathbf{y}^*}{\mathbf{y}^{*\top}\mathbf{U}_b^*\left[\frac{1}{\kappa}\left(\mathbf{I}_{n-q}-\mathbf{D}_b^*\left(\mathbf{D}_b^{*2}+\kappa\mathbf{I}_{n-q}\right)^{-1}\mathbf{D}_b^*\right)\right]\mathbf{U}_b^{*\top}\mathbf{y}^*}\right) \\ =& (n-q)\log\kappa-\sum_{i=1}^{n-q}\log(\gamma_i+\kappa)-(n-q)\log\left(1-\frac{\mathbf{y}^{*\top}\mathbf{U}_b^*\mathbf{D}_b^*\left(\mathbf{D}_b^{*2}+\kappa\mathbf{I}_{n-q}\right)^{-1}\mathbf{D}_b^*\mathbf{U}_b^{*\top}\mathbf{y}^*}{\mathbf{y}^{*\top}\mathbf{y}^*}\right).\end{aligned}$$

The main part of $2\log B_{10}^{\kappa}$ is

$$T_n^{\kappa}=\frac{\mathbf{y}^{*\top}\mathbf{U}_b^*\mathbf{D}_b^*\left(\mathbf{D}_b^{*2}+\kappa\mathbf{I}_{n-q}\right)^{-1}\mathbf{D}_b^*\mathbf{U}_b^{*\top}\mathbf{y}^*}{\mathbf{y}^{*\top}\mathbf{y}^*}.$$

A large value of $T_n^{\kappa}$ supports the alternative hypothesis. Under the null hypothesis,

$$\operatorname{E}T_n^{\kappa}=\frac{1}{n-q}\operatorname{tr}\left(\mathbf{D}_b^{*2}(\mathbf{D}_b^{*2}+\kappa\mathbf{I}_{n-q})^{-1}\right).$$

Under the alternative hypothesis, consider $\boldsymbol{\beta}_1 = c\boldsymbol{\beta}_1^\dagger$ where $\boldsymbol{\beta}_1^\dagger \neq 0$ is a fixed direction and $c > 0$. As $c \to \infty$, $T_n^\kappa \to$. We say $T_n^\kappa$ is consistent along the direction $\boldsymbol{\beta}_1^\dagger$ if. It turns out that, $T_n^\kappa$ is consistent along certain directions of $\boldsymbol{\beta}_1$, while it is inconsistent along other directions. Define $B_{10}^0 = \lim_{\kappa\to 0} B_{10}^\kappa$. Then

$$2\log B_{10}^0 = -\sum_{i=1}^{n-q}\log(\gamma_i) + (n-q)\log\left(\frac{\mathbf{y}^{*\top}\mathbf{y}^*}{\mathbf{y}^{*\top}(\mathbf{X}_b^*\mathbf{X}_b^{*\top})^{-1}\mathbf{y}^*}\right).$$

## 2 Distribution under the null hypothesis

Under the null hypothesis, the distribution of $2\log B_{10}$ does not rely on unknown parameters. Further more, its distribution is valid as long as the distribution of $\epsilon$ is spherically symmetric.

**Proposition 2.** *Under the null hypothesis,*

$$T_n := \frac{\mathbf{y}\top(\mathbf{I}_n - \mathbf{P}_a)\mathbf{X}_b\left(\mathbf{X}_b^\top(\mathbf{I}_n - \mathbf{P}_a)\mathbf{X}_b + \kappa\mathbf{I}_p\right)^{-1}\mathbf{X}_b^\top(\mathbf{I}_n - \mathbf{P}_a)\mathbf{y}}{\mathbf{y}\top(\mathbf{I}_n - \mathbf{P}_a)\mathbf{y}} \sim \frac{\sum_{i=1}^{n-q}\frac{\gamma_i}{\gamma_i+\kappa}Z_i^2}{\sum_{i=1}^{n-q}Z_i^2},$$

*where $\gamma_i$ is the ith eigenvalue of $\mathbf{X}_b^\top(\mathbf{I}_n - \mathbf{P}_a)\mathbf{X}_b$, $i = 1,\dots,n-q$, and $Z_1,\dots,Z_{n-q}$ are iid $\mathcal{N}(0,1)$ random variables.*

Let $\nu_i = \gamma_i/(\gamma_i+\kappa)$, $\bar\nu = (n-q)^{-1}\sum_{i=1}^{n-q}\nu_i$.

**Lemma 1.** *Under the null hypothesis, a necessary and sufficient condition for*

$$\frac{n-q}{\sqrt{2\sum_{i=1}^{n-q}(\nu_i-\bar\nu)^2}}(T_n - \bar\nu) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \tag{1}$$

*is that*

$$\frac{\max_{i\in\{1,\dots,n-q\}}(\nu_i-\bar\nu)^2}{\sum_{i=1}^{n-q}(\nu_i-\bar\nu)^2} \to 0. \tag{2}$$

*Proof.* Note that

$$\frac{n-q}{\sqrt{2\sum_{i=1}^{n-q}(\nu_i-\bar\nu)^2}}(T_n - \bar\nu) \sim \frac{n-q}{\sum_{i=1}^{n-q}Z_i^2}\frac{\sum_{i=1}^{n-q}(\nu_i-\bar\nu)Z_i^2}{\sqrt{2\sum_{i=1}^{n-q}(\nu_i-\bar\nu)^2}}.$$

By Slutsky's theorem, (1) holds if and only if

$$\frac{\sum_{i=1}^{n-q}(\nu_i-\bar\nu)Z_i^2}{\sqrt{2\sum_{i=1}^{n-q}(\nu_i-\bar\nu)^2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

From Lemma 1 of Wang and Xu (2018), (2) is a necessary and sufficient condition for this to hold. □

## 3 Distribution under the alternative hypothesis

# Appendices

## Appendix A haha1

***Proof of Proposition 1.*** We assume $0 < \alpha < 1$ since the case $\alpha = 0$ or 1 is trivial. Note that the condition implies $\int [\varphi(\mathbf{y}) - \alpha] \mathcal{N}_n(0, \phi^{-1}\mathbf{I}_n)(d\mathbf{y}) = 0$. Hence it suffices to prove $\varphi(\mathbf{y}) \geq \alpha$, a.s. We prove this by contradiction. Suppose $\lambda(\{\mathbf{y} : \varphi(\mathbf{y}) < \alpha\}) > 0$. Then there exists a $\eta > 0$, such that $\lambda(\{\mathbf{y} : \varphi(\mathbf{y}) < \alpha - \eta\}) > 0$. We denote $E = \{\mathbf{y} : \varphi(\mathbf{y}) < \alpha - \eta\}$. From Lebesgue density theorem (Cohn, 2013, Corollary 6.2.6), there exists a point $z \in E$, such that, for each $\epsilon > 0$ there is a $\delta_\epsilon > 0$ such that

$$\left| \frac{\lambda(E^\complement \cap C_\epsilon)}{\lambda(C_\epsilon)} \right| < \epsilon,$$

where $C_\epsilon = \prod_{i=1}^n [z_i - \delta_\epsilon, z_i + \delta_\epsilon]$. We put

$$\epsilon = \left( \frac{\sqrt{\pi}}{\sqrt{2}\Phi^{-1}\left(1 - \frac{\eta}{6n}\right)} \right)^n \frac{\eta}{3}.$$

Then for any $\phi > 0$,

$$\begin{aligned}
\alpha \leq & \int_{\mathbb{R}^n} \varphi(\mathbf{y}) \mathcal{N}_n(z, \phi^{-1}\mathbf{I}_n)(d\mathbf{y}) \\
= & \int_{E \cap C_\epsilon} \varphi(\mathbf{y}) \mathcal{N}_n(z, \phi^{-1}\mathbf{I}_n)(d\mathbf{y}) + \int_{E^\complement \cap C_\epsilon} \varphi(\mathbf{y}) \mathcal{N}_n(z, \phi^{-1}\mathbf{I}_n)(d\mathbf{y}) + \int_{C_\epsilon^\complement} \varphi(\mathbf{y}) \mathcal{N}_n(z, \phi^{-1}\mathbf{I}_n)(d\mathbf{y}) \\
\leq & \alpha - \eta + \int_{E^\complement \cap C_\epsilon} \mathcal{N}_n(z, \phi^{-1}\mathbf{I}_n)(d\mathbf{y}) + \int_{C_\epsilon^\complement} \mathcal{N}_n(z, \phi^{-1}\mathbf{I}_n)(d\mathbf{y}) \\
\leq & \alpha - \eta + \left( \frac{\phi}{2\pi} \right)^{n/2} \lambda(E^\complement \cap C_\epsilon) + 2n \left( 1 - \Phi(\sqrt{\phi}\delta_\epsilon) \right) \\
\leq & \alpha - \eta + \left( \frac{\phi}{2\pi} \right)^{n/2} \epsilon (2\delta_\epsilon)^n + 2n \left( 1 - \Phi(\sqrt{\phi}\delta_\epsilon) \right) \\
= & \alpha - \eta + \left( \frac{\sqrt{\phi}\delta_\epsilon}{\Phi^{-1}\left(1 - \frac{\eta}{6n}\right)} \right)^n \frac{\eta}{3} + 2n \left( 1 - \Phi(\sqrt{\phi}\delta_\epsilon) \right).
\end{aligned}$$

Putting

$$\phi = \left( \frac{\Phi^{-1}\left(1 - \frac{\eta}{6n}\right)}{\delta_\epsilon} \right)^2$$

yields the contradiction $\alpha \leq \alpha - (2/3)\eta$. This completes the proof.

$\square$

## Appendix B haha2

## References


Cohn, D. L. (2013). *Measure Theory*. Birkhauser Advanced Texts Basler Lehrbucher. Birkhuser Basel, 2 edition.

Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934.

Maruyama, Y. and George, E. I. (2011). Fully bayes factors with a generalized g -prior. *Ann. Statist.*, 39(5):2740–2765.

Shang, Z. and Clayton, M. K. (2011). Consistency of bayesian linear model selection with a growing number of parameters. *Journal of Statistical Planning and Inference*, 141(11):3463–3474.

Wang, R. and Xu, X. (2018). On two-sample mean tests under spiked covariances. *Journal of Multivariate Analysis*, 167:225 – 249.